ACM Multimedia 2024
Melbourne, Australia

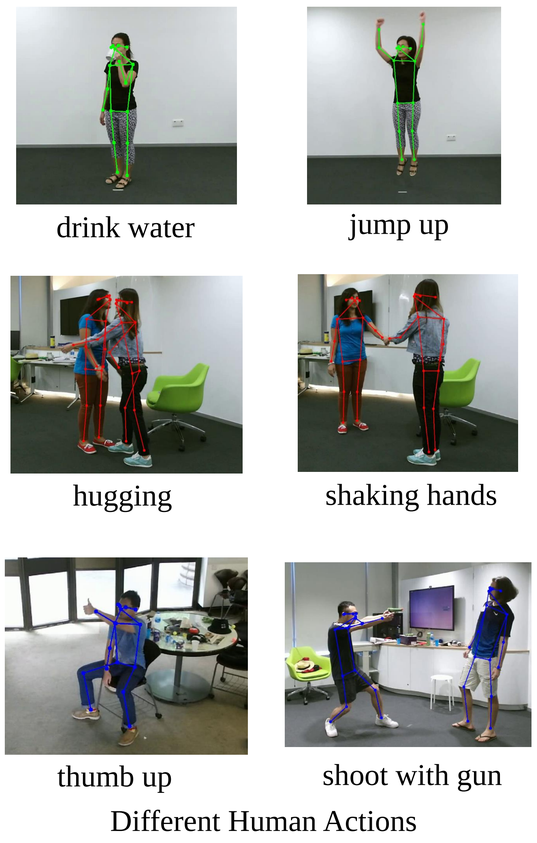
Multi-Modality Co-Learning for Efficient Skeleton-based Action Recognition

*Jinfu Liu, Chen Chen, Mengyuan Liu**

State Key Laboratory of General Artificial Intelligence, Peking University
Center for Research in Computer Vision, University of Central Florida

Abstract

Skeleton-based action recognition has garnered significant attention due to the utilization of concise and resilient skeletons. Nevertheless, the absence of detailed body information in skeletons restricts performance, while other multimodal methods require substantial inference resources and are inefficient when using multimodal data during both training and inference stages. To address this and fully harness the complementary multimodal features, we propose a novel multi-modality co-learning (MMCL) framework by leveraging the multimodal large language models (LLMs) as auxiliary networks for efficient skeleton-based action recognition, which engages in multi-modality co-learning during the training stage and keeps efficiency by employing only concise skeletons in inference. Our MMCL framework primarily consists of two modules. First, the Feature Alignment Module (FAM) extracts rich RGB features from video frames and aligns them with global skeleton features via contrastive learning. Second, the Feature Refinement Module (FRM) uses RGB images with temporal information and text instruction to generate instructive features based on the powerful generalization of multimodal LLMs. These instructive text features will further refine the classification scores and the refined scores will enhance the model's robustness and generalization in a manner similar to soft labels. Extensive experiments on NTU RGB+D, NTU RGB+D 120 and Northwestern-UCLA benchmarks consistently verify the effectiveness of our MMCL, which outperforms the existing skeleton-based action recognition methods. Meanwhile, experiments on UTD-MHAD and SYSU-Action datasets demonstrate the commendable generalization of our MMCL in zero-shot and domain-adaptive action recognition.

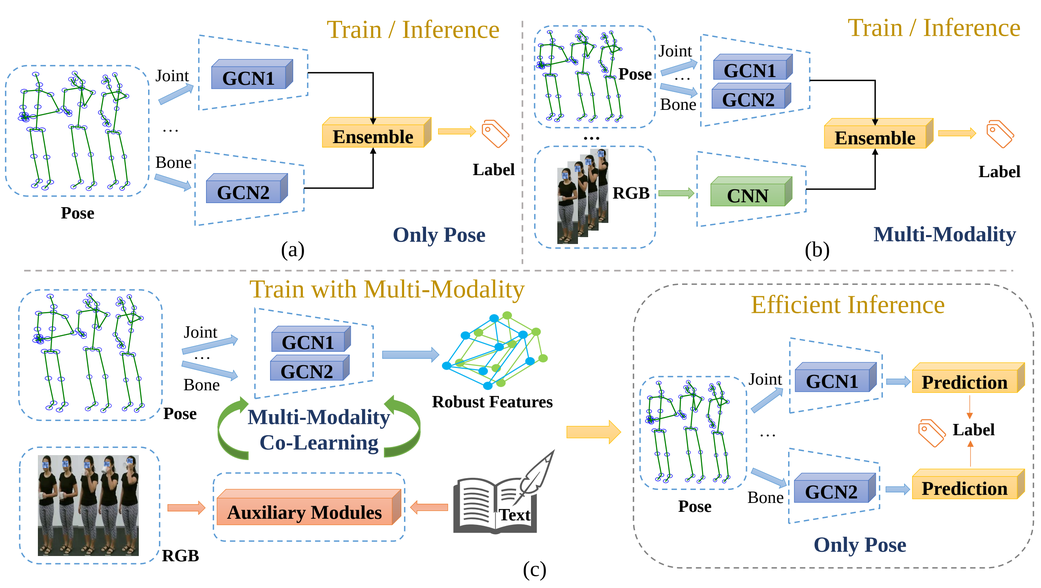


Github | Arxiv | Paper With Code

Contributions

- We propose a novel multi-modality co-learning (MMCL) framework for efficient skeleton-based action recognition, which empowers mainstream GCN models to produce more robust and generalized feature representations by introducing multi-modality co-learning during the training stage, while maintain efficiency by only using concise skeletons in inference.
- Our proposed MMCL framework is the first to introduce multimodal LLMs for multi-modality co-learning in skeleton-based action recognition. Meanwhile, our MMCL is orthogonal to the backbones and thus can be applied to optimize mainstream GCN models by using different multimodal LLMs. Due to the generalization of multimodal LLMs, our MMCL can be transferred to domain-adaptive and zero-shot action recognition.
- Extensive experiments on three popular benchmarks namely NTU RGB+D, NTU RGB+D 120 and Northwestern-UCLA datasets verify the effect of our MMCL framework by outperforming existing skeleton-based methods. Meanwhile, experiments on SYSU-ACTION and UTD-MHAD datasets from different domains indicate that our MMCL exhibits commendable generalization in both domain adaptive and zero-shot action recognition.

Introduction

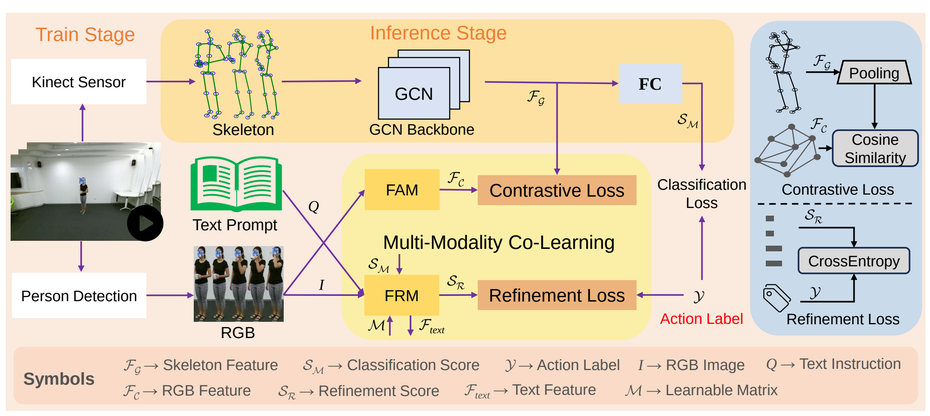


- Human action recognition is an important task in video understanding. The human action conveys central information like body tendencies and thus helps to understand the person in videos. In pursuit of precise action recognition, diverse video modalities have been explored, such as skeleton sequences, RGB images, text descriptions, depth images and human parsing.
- Skeleton-based methods only using skeletons during training and inference stages will restrict recognition performance due to the inherent defects of skeletal modality. For instance, the skeleton modality lacks the ability to depict detailed body information (e.g. appearance and objects) and has difficulty in fine-grained recognition when dealing with similar actions.
- Multimodal-based methods can make up for the defects of a single modality and provide more comprehensive information for fine-grained action classification by leveraging the complementary nature of multimodal data. Nevertheless, they suffer from the drawbacks of requiring significant inference resources and appear less efficient when deployed on edge devices due to the use of multi-modality in both training and inference stages.

A Question?

- ***How to better leverage the complementary nature of multi-modality while retaining the efficient inference with single-skeleton modality?***
- We propose a multi-modality co-learning (MMCL) framework for efficient skeleton-based action recognition, which enhances the model's performance and generalization via multi-modality co-learning, while only using the concise skeleton in the inference stage to preserve the efficiency.

Method



Feature Alignment Module

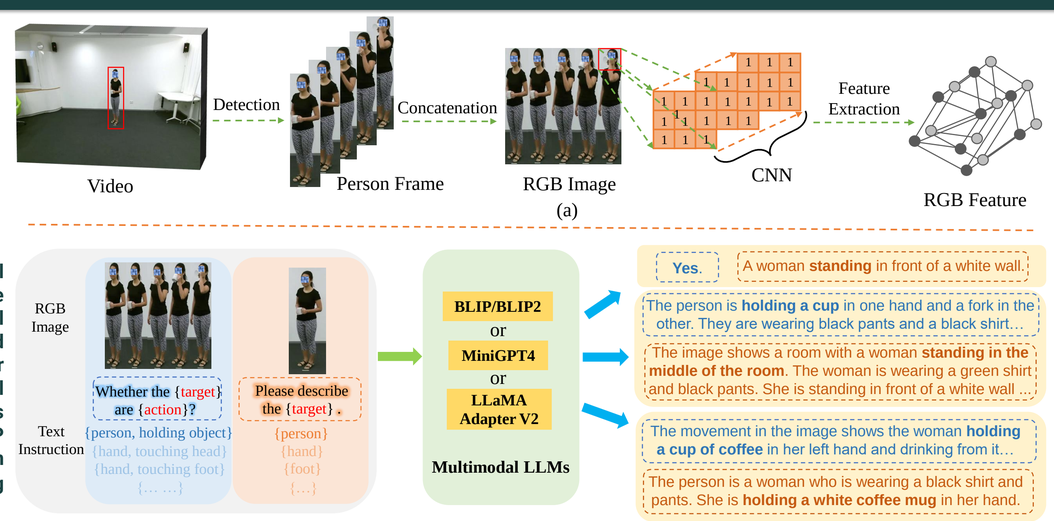
$$F_c = M_{align}\left(\Theta\left(Norm(\Gamma(I))\right)\right)$$

$$L_c = \frac{1}{2N}\sum_{i=1}^{N}\left[L(F_g^i, F_c^i) + L(F_c^i, F_g^i)\right]$$

$$L(F_g^i, F_c^i) = log\frac{e^{\theta(F_g^i, F_c^i)/\tau}}{e^{\theta(F_g^i, F_c^i)/\tau} + Neg(F_g^i, F_c^i)}$$

$$Neg(F_g^i, F_c^i) = \sum_{k\neq i}\left(e^{\theta(F_g^i, F_c^k)/\tau} + e^{\theta(F_g^i, F_g^k)/\tau}\right)$$

The FAM utilizes Θ to model high-level features from RGB image I. Before extracting RGB features, the FAM will resize and normalize the image to speed up the convergence of the model. In order to achieve alignment with the global skeleton features, the RGB features extracted by CNN will go through the MLP operation M_{align}. This alignment operation facilitates subsequent contrastive learning between two features.



GCN Backbone

$$H^{l+1} = \sigma\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H^lW^l\right)$$

The H^l is the joint features at layer l and σ is the activation function. D is the degree matrix of joints and W^l is the learnable parameter of the l_{th} layer. A is the adjacency matrix representing joint connections.

Feature Refinement Module

$$F_{text} = \Phi\left(VQA(I, Q)\right) \qquad S_R = \sum_{i=1}^{N} F_{text}^i M_i S_M$$

The FRM inputs text instruction and RGB image into the LLMs in the form of Visual Question Answering(VQA). Then the token features of the multimodal LLM will be retained. The FRM unifies text features obtained from different input samples into the same dimension through Φ.

Loss Function

$$L_R = -\sum_i^N y^i log S_R^i$$

$$L_{loss} = L_{cls} + \lambda_1 L_c + \lambda_2 L_R$$

We use Cross-Entropy (CE) loss as the classification loss. Our MMCL uses contrastive loss and refinement loss as the auxiliary loss to constrain the training of the whole network, thereby introducing multimodal features into skeleton-based action recognition.

Data Processing

$$I = \oplus_m U_n^m\left(\Gamma\left(\varepsilon(k_1)\right), \Gamma\left(\varepsilon(k_2)\right), \ldots, \Gamma\left(\varepsilon(k_n)\right)\right)$$

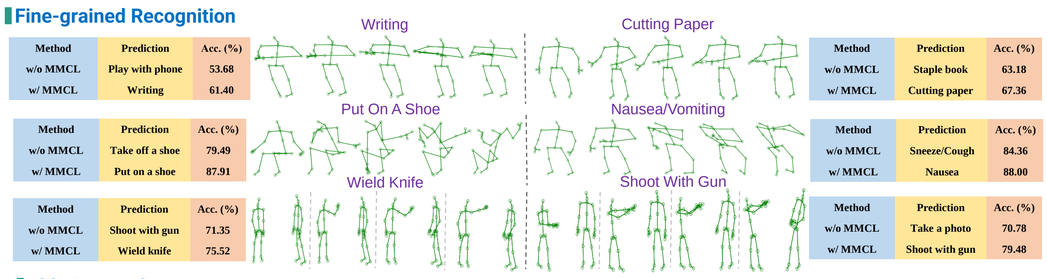
We select m frames from a action video of drinking water and concatenate them from left to right in chronological order, thus forming an RGB image containing the temporal information of drinking water.

Experiments

State-of-the-art Result

Method	NTU60(%)		NTU120(%)		NW-UCLA(%)
	X-Sub	X-View	X-Sub	X-Set	
Shift-GCN	90.7	96.5	85.9	87.6	94.6
DynamicGCN	91.5	96.0	87.3	88.6	-
MS-G3D	91.5	96.2	86.9	88.4	-
MST-GCN	91.5	96.6	87.5	88.8	-
MG-GCN	92.0	96.6	88.2	89.3	-
CTR-GCN	92.4	96.8	88.9	90.6	96.5
PSUMNet	92.9	96.7	89.4	90.6	-
ACFL-CTR	92.5	97.1	89.7	90.9	-
SAP-CTR	93.0	96.8	89.5	91.1	-
InfoGCN	93.0	97.1	89.8	91.2	97.0
FR-Head	92.8	96.8	89.5	90.9	96.8
SkeletonGCL	93.1	97.0	89.5	91.0	96.8
Koopman	92.9	96.8	90.0	91.3	97.0
VPN	93.5	96.2	86.3	87.8	93.5
TSMF	92.5	97.4	87.0	89.1	-
DRDIS	91.1	94.3	81.3	83.4	-
LST	92.9	97.0	89.9	91.1	97.2
MMCL	**93.5**	**97.4**	**90.3**	**91.7**	**97.5**

Fine-grained Recognition



Method	Prediction	Acc. (%)
w/o MMCL	Play with phone	53.68
w/ MMCL	Writing	61.40

Method	Prediction	Acc. (%)
w/o MMCL	Staple book	63.18
w/ MMCL	Cutting paper	67.36

Method	Prediction	Acc. (%)
w/o MMCL	Take off a shoe	79.49
w/ MMCL	Put on a shoe	87.91

Method	Prediction	Acc. (%)
w/o MMCL	Sneeze/Cough	84.36
w/ MMCL	Nausea	88.00

Method	Prediction	Acc. (%)
w/o MMCL	Shoot with gun	71.35
w/ MMCL	Wield knife	75.52

Method	Prediction	Acc. (%)
w/o MMCL	Take a photo	70.78
w/ MMCL	Shoot with gun	79.48

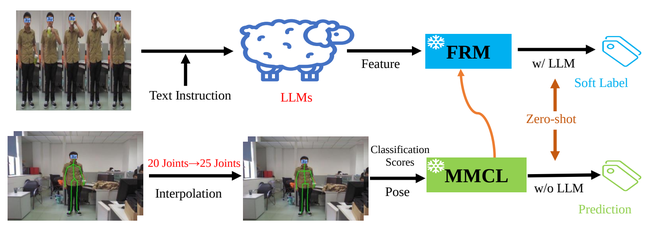
Ablation Studies

Methods	Modality	Acc. (%)
CTRS-GCN w/o MMCL	joint	83.88
CTRS-GCN w/ MMCL	**joint**	**84.84**[↑0.96]
CTR-GCN w/o MMCL	joint	85.01
CTR-GCN w/ MMCL	**joint**	**85.79**[↑0.78]
CTR-GCN w/o MMCL	bone	86.34
CTR-GCN w/ MMCL	**bone**	**87.32**[↑0.98]
CTR-GCN w/o MMCL	joint motion	81.23
CTR-GCN w/ MMCL	**joint motion**	**83.19**[↑1.96]
CTR-GCN w/o MMCL	bone motion	81.66
CTR-GCN w/ MMCL	**bone motion**	**82.92**[↑1.26]

Methods	CNN	Acc. (%)
CTRS-GCN w/o FAM	-	83.88
CTRS-GCN w/ FAM	ResNet34	84.44[↑0.56]
CTRS-GCN w/ FAM	**Inception-V3**	**84.48**[↑0.60]
CTR-GCN w/o FAM	-	85.01
CTR-GCN w/ FAM	ResNet34	85.57[↑0.56]
CTR-GCN w/ FAM	**Inception-V3**	**85.62**[↑0.61]

Methods	LLMs	Acc. (%)
CTRS-GCN w/o FRM	-	83.88
CTRS-GCN w/ FRM	**BLIP**	**84.10**[↑0.22]
CTR-GCN w/o FRM	-	85.01
CTR-GCN w/ FRM	MiniGPT-4	85.42[↑0.41]
CTR-GCN w/ FRM	**BLIP**	**85.60**[↑0.59]

Zero-shot and Domain-adaptive



Method	Top-1/Top-5 Acc. (%)	
	UTD-MHAD	SYSU-Action
CTRS-GCN (J)	37.70/69.63	37.50/53.33
CTR-GCN (J)	48.17/74.87	27.50/51.67
HD-GCN (J CoM-1)	46.07/74.87	26.27/58.33
Ours (w/o BLIP Refine)	**52.88/81.16**	**39.17/76.67**
Ours (w/ BLIP Refine)	**54.97/84.29**	**42.50/80.83**

Acknowledgments

This work was supported by Natural Science Foundation of Shenzhen (No. JCYJ20230807120801002), National Natural Science Foundation of China (No. 62203476).

https://github.com/liujf69/MMCL-Action